On Calibrated Predictions for Auction Selection Mechanisms

H. Brendan McMahan
mcmahan@google.com

Omkar Muralidharan
omuralidharan@google.com

Google, Inc.

Abstract

Calibration is a basic property for prediction systems, and algorithms for achieving it are well-studied in both statistics and machine learning. In many applications, however, the predictions are used to make decisions that select which observations are made. This makes calibration difficult, as adjusting predictions to achieve calibration changes future data. We focus on click-through-rate (CTR) prediction for search ad auctions. Here, CTR predictions are used by an auction that determines which ads are shown, and we want to maximize the value generated by the auction.

We show that certain natural notions of calibration can be impossible to achieve, depending on the details of the auction. We also show that it can be impossible to maximize auction efficiency while using calibrated predictions. Finally, we give conditions under which calibration is achievable and simultaneously maximizes auction efficiency: roughly speaking, bids and queries must not contain information about CTRs that is not already captured by the predictions.

1 Introduction

Calibration is a fundamental measure of accuracy in prediction problems: if we group all the events a predictor says happen with probability p, about a p fraction should occur. This property has been extensively studied in the stochastic and online settings.

We study problems where the predictions themselves partially determine which events occur. Our general approach applies to many problems where predictions are used to make decisions, but we are motivated in particular by the application to search engine advertising. Over the past decade, this business has grown to tens of billions of dollars, and prediction systems play a fundamental role.

In a typical interaction, first a user does a query (say "flowers") on a search engine. Then, the search engine selects a set of candidate ads that can be shown on the given query, based on keywords provided by advertisers. These

components can be reasonably approximated by an IID process. A prediction is made for each candidate ad, and an auction ranks the ads based on the prediction and the bid of the advertiser. Typically, the bid indicates the value of a click to the advertiser, and the score is simply the product of the bid and the prediction, giving an estimate of the value generated by showing the ad. Finally, some of these ads are shown to the user (we consider two models: the single top-ranked ad is shown, or all the ads with scores above a certain threshold are shown). This auction selection mechanism has been extensively studied, and has many nice properties [19, 8].

In this setting, an important measure of the quality of the predictions is how much value the auction generates (equivalently, how *efficient* are the allocations produced by the auction). The auction mechanisms we consider are in fact designed to maximize the combined value to the search engine and advertiser if bids accurately reflect value and the true click-through-rates (CTRs) are known.

The algorithm used to predict CTRs for such a system faces many constraints already, for example, the need to process enormous volumes of data quickly and produce predictions with extremely low latency (e.g., [13]). Thus, rather than advocating new algorithms, we focus on applying a post-correction via a prediction map to the outputs of an existing system in order to improve the quality of the predictions.

We consider two main questions. Informally stated: 1) Do efficiency-maximizing prediction maps with calibration properties exist, and can they can be found computationally efficiently? 2) If we iteratively calibrate our predictions so they match observed CTRs, does the process converge? And if so, is this prediction map efficiency maximizing?

Outline and Summary of Results We formalize our model and questions in Section 2, where we introduce two primary variants of the selection mechanism that lead to different properties; Section 3 and 4 investigate these mechanisms in the general case. We demonstrate that without further assumptions, in both our models it may be impossible for a deterministic prediction map to produce calibrated predictions on the ads it serves, and iterative calibration procedures can fail badly. Since some deterministic map always maximizes value, this is unfortunate. When all ads above a certain threshold are shown, we give an algorithm for finding this value-maximizing map in polynomial time, but when the single highest-rated ad is shown, we prove finding the value-maximizing map is NP-hard (even if we knew the true CTRs).

In Section 5 we introduce additional assumptions that are sufficient to guarantee calibration procedures are well-behaved. While these assumptions are fairly strong, they are not unreasonable for real systems. Our strongest assumption is essentially that in all cases bid and query provide no more information than the raw prediction about average CTRs; under this assumption, we can show in both selection models a value-maximizing and calibrated prediction map exists. Under threshold selection, somewhat weaker conditions are in fact sufficient.

Related Work Calibration has been extensively studied. Much of the earliest work is in the probabilistic forecasting literature [1, 6, 18]. Calibration is particularly important when comparing predictors, since two sets of calibrated predictions can be fairly evaluated by how concentrated they are on observed outcomes [7, 12, 11]. Calibration also makes it easier to use predictions. For example, it is easier to threshold the output of a calibrated classifier to minimize weighted classification error [5].

Not all prediction systems are naturally calibrated. However, when examples are drawn IID, if we have a good but uncalibrated predictor, we can calibrate it by applying a prediction map. For example, boosted trees are uncalibrated, but become excellent probability estimators after calibration [16, 2]. The two most common methods for calibration are Platt scaling, which is equivalent to logistic regression, and isotonic regression [17, 20, 15, 3].

Calibration is also studied in the online setting, where no stochastic assumptions are made on the sequence of examples; in the worst case, they could be chosen by an adversary that sees our predictions. It is easy to see that in this setting, no deterministic classifier (or prediction map) can produce calibrated predictions for all sequences. However, if the system is allowed to use randomness (that is, predict a distribution), then calibration can be achieved ([9, 10] and [4, Sec 4.5]).

2 Problem Formalization

The interaction of calibration and selection has received little direct attention in the literature, so constructing a suitable model requires some care: we require a formulation that is theoretically tractable but still captures the key characteristics of the real-world problems of interest.

We begin by defining our units of prediction (queries and ads) and the mechanism used to select them (auctions). We assume a fixed, existing prediction system provides a raw prediction for each ad; our study will then concern prediction maps, functions that attempt to map these raw predictions to calibrated probabilities. Once this framework is established, we can formally state the questions we study.

We model the interaction between a search engine's users and advertising system. There is a fixed finite set of queries $\mathcal{Q}$ (strings like "flowers" or "car insurance" typed into the search engine), which are chosen according to distribution $\Pr^{\mathcal{Q}}(q)$ for $q \in \mathcal{Q}$. There is also a fixed finite set of ads $\mathcal{C}$ which can be shown alongside queries. Each ad $i \in \mathcal{C}$ is defined by tuple (p_i, b_i, z_i, q_i) where $q_i \in \mathcal{Q}$ is the (only) query for which ad i can show,[1] p_i is the true probability of a click, b_i is the bid (the maximum amount the advertiser is willing to pay for a click), and $z_i \in \{1, \ldots, K\}$ is a bucketed estimate of p_i (we call z_i the raw prediction). That is, we assume the predictions of the underlying prediction system have been discretized into K buckets. We drop the q (and sometimes z)

[1] This is without loss of generality, as we can always replicate ads for each query to which the advertiser has targeted the ad.

from the ad tuples when those values are clear from context. Each ad can show for a single query q, so we define $\mathcal{C}(q) \equiv \{i \,|\, q_i = q\}$, the indexes of the candidate ads for query q.

Our goal is to find good prediction maps $f : \{1, \ldots, K\} \to [0, 1]$. The prediction map will be used in the auction selection mechanism: First, a query is sampled from $\Pr^{\mathcal{Q}}$, and then the candidate ads for that query are ranked by $b \cdot f(z)$ (we drop the subscripts when we mean an arbitrary ad). We consider two models for which ads show:

`ONE`: We only show a single ad. If multiple ads achieve the highest value of $b \cdot f(z)$, we pick one uniformly at random.

`ALL`: We show all ads where $b \cdot f(z) - 1 > 0$.

Mechanism `ONE` models the case of an oversold auction, where ads with different raw predictions z must compete for a single position. Mechanism `ALL` models the case where all eligible ads with positive predicted value can be shown. In general, mechanism `ALL` is much easier to work with theoretically, because for $z_1 \neq z_2$, changing $f(z_1)$ does not change which ads with prediction z_2 are shown. In either case, we assume any candidate (p, b, z) which is shown is clicked with probability p.[2]

Distributions on Ads Other than the distribution $\Pr^{\mathcal{Q}}$, all probabilities and expectations will be with respect to some distribution on the set of candidate ads $\mathcal{C}$. Two distributions will be of particular importance: $\Pr_{\mathcal{C}}$, the uniform distribution over candidate ads, and $\Pr_f$, the distribution of ads shown by a prediction map f. We formalize these as follows:

$\Pr_{\mathcal{C}}$ is the distribution on ads where $\Pr_{\mathcal{C}}(i)$ is proportional to $\Pr^{\mathcal{Q}}(q_i)$. That is, letting $C \equiv \sum_{i \in \mathcal{C}} \Pr^{\mathcal{Q}}(q_i)$, we have $\Pr_{\mathcal{C}}(i) = \frac{\Pr^{\mathcal{Q}}(q_i)}{C}$. This is not the same as choosing a random query q from $\Pr^{\mathcal{Q}}$ and then choosing a random candidate. For example, suppose there are two queries q_1 and q_2, with $\Pr^{\mathcal{Q}}(q_1) = \frac{1}{2}$ and $\Pr^{\mathcal{Q}}(q_2) = \frac{1}{2}$. There is one candidate a_1 for query q_1, and two candidates, a_2 and a_3 for query q_2. Then, $\Pr_{\mathcal{C}}(a_i) = 1/3$ for each ad, which means the marginal probability $\Pr_{\mathcal{C}}(q_1) = \frac{1}{3}$ and $\Pr_{\mathcal{C}}(q_2) = \frac{2}{3}$. One can think of $\Pr_{\mathcal{C}}$ as the distribution on ads shown if we showed all the eligible candidates for each query that occurs.

$\Pr_f$ for a prediction map f is the distribution on ads where $\Pr_f(i)$ is proportional to $w_i \equiv \Pr^{\mathcal{Q}}(q_i)\Pr(\text{ad } i \text{ shows} \,|\, q_i, f)$. The second term is actually only random in the case of selection mechanism `ONE`, when randomness is used to break ties. The distribution $\Pr_f$ is thus the distribution on ads shown when serving using prediction map f. Using this notation, $\Pr_f(i \,|\, q) = \Pr(\text{ad } i \text{ shows} \,|\, q_i, f)$.

We use $\mathbb{E}_{\mathcal{C}}[\cdot]$ and $\mathbb{E}_f[\cdot]$ for the corresponding expectations.

[2]This ignores the well-known issue of position normalization; this aspect of the problem is largely orthogonal to our work.

Calibration We say a prediction map f is calibrated on a distribution on ads D if

$$\forall z, \quad \underbrace{\mathbb{E}_{(p,b,z,q)\sim D}[p \,|\, z]}_{\text{Average CTR given } z} = \underbrace{f(z)}_{\text{Predicted CTR given } z}.$$

The choice of the distribution D in the above definition is critical; a single f will in general not be able to achieve calibration for multiple D. For the auction selection problem, the natural distribution to consider is $\Pr_f$. Thus, we will be particularly concerned with finding *self-calibrated* prediction maps f, which satisfy

$$\forall z, \quad \mathbb{E}_f[p \,|\, z] = f(z).$$

In general one may not be able to estimate $\mathbb{E}_f[\, p \mid z \,]$ exactly, and so calibration will only be approximately achievable. This issue is orthogonal to our results, so we assume that the necessary expected quantities can be estimated exactly. Thus, we emphasize that our negative results are a fundamental limitation, rather than a byproduct of insufficient data.

Auction Efficiency In addition to calibration, we are concerned with how the choice of f impacts the auction mechanism. The expected value of showing ad (p, b) is $p \cdot b - \text{cost}$, where we take cost $= 1$ for selection mechanism `ALL`, and cost $= 0$ for `ONE`. We assume the bid b reflects the true value to the advertiser of a click, which is justified by the incentives of the auction under a suitable pricing scheme [19]. The cost can be viewed as the cost per impression of showing the ad (either a cost incurred by the user doing the query or incurred by the search engine itself). In practice such costs might be different for clicked versus unclicked ad impressions, and might vary depending on the ad and query. Extending our results to such a models would add a significant notational burden, so we focus on the simplest interesting cost models.

For a given query q, the expected value generated is

$$\sum_{i \in C(q)} \Pr(\text{ad } i \text{ shows} \,|\, f, q)(p_i b_i - \text{cost}).$$

The expected value per query is just

$$\begin{aligned} \text{EV}(f) &= \sum_{q \in Q} \Pr^{\mathcal{Q}}(q) \sum_{i \in C(q)} \Pr_f(i \,|\, q)(p_i b_i - \text{cost}) \\ &= \sum_{i \in C} w_i (p_i b_i - \text{cost}). \end{aligned}$$

We say an $f^* \in \arg\max_f \text{EV}(f)$ is *efficiency maximizing*. Our goal is to find an f that transforms the z into the best possible predictions in terms of efficiency. Note that if it was possible to predict exactly p_i for ad i, these predictions would maximize efficiency.

Questions Ideally, we would like to use prediction maps that are self-calibrated and efficiency-maximizing; we say such prediction maps are *nice*, and say a problem instance is *nice* if such a map exists.

First, we consider questions relating to the offline problem where we have access to all the problem data. Note that there must exist an efficiency-maximizing prediction map.[3]

Q1 Are all problem instances nice? That is, do self-calibrated efficiency-maximizing prediction maps always exist?

Q2 Can an efficiency-maximizing prediction map, even one that is not self-calibrated, be found in polynomial time?

In practice, we are further concerned with learning a good prediction map from observed data. Suppose we start with some f_0, for example the function that gives the predictions of the underlying system. Then, we serve some large number of queries with this f_0, and observe the results. We would like to then train an improved f_1 from this data, serve another large batch of queries ranked using f_1, then train an f_2, etc.

A natural procedure is to choose f_t so that the predictions on the ads shown in batch $t-1$ would have been calibrated under f_t. Of course, when we then select ads using f_t on the next batch, we may show different ads. Formally, define $T : [0,1]^K \to [0,1]^K$ (a function from prediction maps to prediction maps) by $T(f) = f'$ where

$$f'(z) = \begin{cases} \mathbb{E}_f[p \,|\, z] & \text{when } \Pr_f(z) > 0 \\ f(z) & \text{otherwise.} \end{cases}$$

We assume we have enough data in each batch so that we can calculate $\mathbb{E}_{f_{t-1}}[p \,|\, z]$ exactly. Then, we ask:

Q3 Does T always have at most a small (polynomial) number of fixed points?

Q4 Does T always have at least one fixed point where ads are shown?

Q3 is important, because with an affirmative answer we could potentially enumerate the fixed points and find the best one from an efficiency perspective. A negative answer to Q4 implies the iterative calibration procedure will cycle. To see this, note that for a given starting point f_0, subsequent $f_t(z)$ can only take on finitely many values: $\mathbb{E}[p \,|\, z]$ for some distribution of ads that show (finitely many values), or $f_0(z)$. That means that T maps some finite set of calibration maps into itself. Since it has no fixed points, T is a permutation and so must cycle.

In the next two sections, we address these questions in the general case (putting no additional restrictions on the problem instances).

[3]Note EV depends only on the ordering of the ads for each query induced by f, and so over all possible f, EV takes on only a finite number of distinct values.

Ad	CTR	bid	min $\hat{p}$	EV	cumulative CTR
1	0.1	$1/(0.1) = 10.0$	0.10	0.00	0.10
2	0.2	$2/(0.1 + 0.2) \approx 6.7$	0.15	0.33	0.15
3	0.3	$3/(0.1 + 0.2 + 0.3) = 5.0$	0.20	0.50	0.20
4	0.4	$4/(0.1 + \cdots + 0.4) = 4.0$	0.25	0.60	0.25

Figure 1: An example with 5 fixed points, one for each prefix of the list of ads. For each i, setting $\hat{p}$ to the value in the "min $\hat{p}$" column induces a fixed point where ads $1, \ldots, i$ show. The fifth fixed point is the degenerate one that shows no ads, with say $\hat{p} = 0$.

3 Mechanism ALL: Threshold Selection

In this section, we consider the case where we select ads by mechanism ALL, that is, we show all ads where $b \cdot f(z) - 1 \geq 0$.

We will show that an efficiency-maximizing prediction map can be found efficiently (Q2), but without further assumptions, Q1, Q3, and Q4 are answered in the negative. We prove the negative results first; for this purpose, it is sufficient to construct counter-examples.

In this section, the examples we construct all require only a single query where all of the candidates have the same raw prediction z. Thus, choosing prediction map reduces to choosing a single value $\hat{p} \in [0, 1]$. The selection rule simply shows all candidates where $b \cdot f(z) = b \cdot \hat{p} \geq 1$.

Q1: All fixed points can have bad efficiency Consider an example with $2n + 1$ candidate ads, divided into three classes, with ads given as (p, b) tuples:

A) 1 ad is $(0.5, 2.0)$, shown if $\hat{p} \geq 0.5$

B) n ads are $(1, 1.9)$, shown if $\hat{p} \geq 1/1.9 \approx 0.53$

C) n ads are $(0, 1.8)$, shown if $\hat{p} \geq 1/1.8 \approx 0.56$

We either show no ads, A, $A+B$, or $A+B+C$. Choosing $\hat{p} = 0.5$ is a fixed point (it only shows the first ad) which generates value $0.5 \cdot 2 - 1 = 0$. Using $\hat{p} = 0.54$ shows $A + B$, and generates value $0.9n$. But, this is not a fixed point: the observed CTR is near one (for large n). Showing all the ads (which occurs for any $\hat{p} > 1/1.8$) is not a fixed point, and generates negative value, since ads from class C generate value $-n$.

Q3: An example with exponentially many fixed points Suppose there are n candidates (p_i, b_i) where the p_i are distinct, and we have indexed by i so that p_i is strictly increasing. Further, suppose $b_i = \frac{i}{p_{1:i}}$, a decreasing sequence (using the shorthand $p_{1:i} \equiv \sum_{j=1}^{i} p_j$). Pick any $i \in \{1, \ldots, n\}$, and let $\hat{p} = \frac{1}{b_i}$. We show candidate j if $b_j \hat{p} = \frac{b_j}{b_i} \geq 1$. Since the bids are decreasing, we show candidate j if and only if $j \leq i$. Thus, serving with $\hat{p} = \frac{1}{b_i} = \frac{p_{1:i}}{i}$ we

show candidates $1, \ldots, i$, and so the average CTR is in fact $\hat{p}$. Thus, for any $i \in \{1, \ldots, n\}$, there is a fixed-point $\hat{p}$ that shows ads $\{1, \ldots, i\}$. Figure 1 shows an example of this construction. If we have m queries each with a distinct fixed raw prediction z and n candidates constructed in this manner, we can choose a per-query fixed point independently for each query, for n^m distinct fixed points.

Q4: An example with no fixed points Consider a single query with two candidates, $(p_1 = 0.7, b_1 = 4, z)$ and $(p_2 = 0.1, b_2 = 2, z)$. For any $\hat{p} \geq 0.5$, both ads show and we observe a click-through-rate of 0.4, so no such $\hat{p}$ can be self-calibrated. For any $\hat{p} \in [0.25, 0.5)$, only ad 1 shows, and we observe a click-through rate of 0.7. For $\hat{p} \in [0, 0.25)$, we don't show any ads. Thus, there is no non-trivial fixed point; assuming we start with $\hat{p} \geq 0.25$, the calibration procedure will cycle between 0.7 and 0.4.

Q2: Calculating the efficiency-maximizing f The above examples show that self-calibrated prediction maps may not exist, and that even if they do, they need not maximize efficiency.

Nevertheless, given access to the full problem data (including true click-through rates) one might be interested in calculating an efficiency maximizing prediction map. The following algorithm accomplishes this in polynomial time.

We define f^* by considering each $z' \in \{1, 2, \ldots, K\}$ independently:

1. Consider the set of candidates (p, b, z, q) where $z = z'$, and sort these candidates in decreasing order of bid, for $j = 1, \ldots, n_j$. We must show some prefix of this list. In particular, if we set $\hat{p} = 1/b_j$ and $b_{j+1} < b_j$, then we will show exactly ads $1, \ldots, j$.

2. For each j where $b_{j+1} < b_j$, compute the expected value per query of using $\hat{p}_j = 1/b_j$ (which shows ads 1, ..., j). This can be computed as

$$\mathrm{EV}(\hat{p}_j) = \sum_{i=1}^{j} \Pr^{\mathcal{Q}}(q_i)(p_i \cdot b_i - 1).$$

3. Let $f(z) = \hat{p}_{j^*}$ where $\hat{p}_{j^*}$ is the value that maximizes $\mathrm{EV}(\hat{p}_j)$.

While this result is interesting theoretically (especially in contrast to results in the next section), we note it is not likely to be useful in practice: if it was possible to estimate p_i accurately for each ad, then one could simply throw out the coarser-grained predictions z_i and use these estimates.

4 Mechanism `ONE`: Selecting One Ad

In this section, we consider results for selection mechanism `ONE`. When there is only a single query, or only a single raw prediction, selection mechanism `ONE` can be quickly analyzed, and our questions are in fact answered in the affirmative, except for Q3. But in non-trivial cases, we again show negative answers to all four questions.

Single query, multiple raw predictions Selection mechanism `ONE` becomes rather degenerate under a single query. We show how to construct a nice f, answering Q1 and Q2, and Q4 in the affirmative.

For each raw prediction $z' \in \{1, \ldots, K\}$, observe that if an ad with $z_i = z'$ shows, it must be an ad that has bid $b(z') \equiv \max_{j:z_j=z'} b_j$. Thus, if an ad with z' shows, the expected value generated is $b(z') \cdot \mathbb{E}_{\mathcal{C}}[p \mid z', b(z')]$, where $\mathbb{E}_{\mathcal{C}}[p \mid z', b(z')]$ is the average click-through-rate of ads with $z = z', b = b(z')$. We can guarantee we obtain this value by simply setting $f(z') = \mathbb{E}_{\mathcal{C}}[p \,|\, z', b(z')]$ and $f(z) = 0$ for all $z \neq z'$. Note that this f is self-calibrated because ties are broken uniformly at random under selection mechanism `ONE`, answering Q4 in the affirmative. We obtain maximum efficiency by using the f that only shows ads with raw prediction

$$z^* = \arg\max_z b(z) \cdot \mathbb{E}_{\mathcal{C}}[p \,|\, z, b(z)].$$

Let f_z be the f function that only shows candidates with the given z value. Thus, f_{z^*} is nice. However, we can define a more satisfying f^* by

$$f^*(z) = \mathbb{E}_{f_z}[p \,|\, z].$$

We only show ads (b, z) where $b \cdot f^*(z)$ achieves the argmax value over candidates, and in fact

$$b \cdot f^*(z) = b(z) \cdot \mathbb{E}_{f_z}[p \,|\, z],$$

and so we still maximize efficiency.

The answer to Q3 is negative: iterative calibration can have exponentially many fixed points. Suppose each ad i has a distinct z_i, and $p_i = b_i^{-1}$. Let $\mathcal{I}$ be any subset of the ads and define $f_{\mathcal{I}}$: $f_{\mathcal{I}}(z_i) = p_i$ for $i \in \mathcal{I}$, $f_{\mathcal{I}}(z_i) = 0$ for $i \notin \mathcal{I}$. Then, under $f_{\mathcal{I}}$ all ads in $\mathcal{I}$ tie, so we show them randomly. Each of the $2^{|\mathcal{C}|}$ subsets of $\mathcal{C}$ thus corresponds to a self-calibrated prediction map that shows a different set of ads.

Multiple queries, single raw prediction Under mechanism `ONE`, if there is a single raw prediction z made for all candidates (on all queries), then the ads that show are in fact independent of the value $\hat{p} = f(z) > 0$: for each query, we always randomly pick one of the candidates with the highest bid. Thus, any $\hat{p} > 0$ is efficiency-maximizing, and we can choose $\hat{p}$ equal to the average observed CTR to obtain self-calibration. Thus, in this case we answer Q1- Q4 in the affirmative.

Q2: NP-hardness in general In general (with at least two distinct raw predictions and at least two queries), under selection mechanism `ONE`, the offline problem of finding the efficiency-maximizing prediction map f is NP-hard, even if all bids are 1. We show this using a reduction from the minimum feedback arc set (MFAS) problem on tournaments (see, for example, Kleinberg et al. [14]).

In this problem, there are n players, $\{1, \ldots, n\}$, that have just completed a tournament where every pair of players has played. The MFAS for this problem

is a ranking of the players that minimizes the number of upsets; that is, if μ_i is the rank of player i, we want a ranking μ that minimizes the number of times $\mu_i > \mu_j$, but player j beat player i.

We encode this problem as an auction efficiency maximization problem as follows: There are n distinct z values, $1, \ldots, n$, one for each player, and there are $\frac{1}{2}n(n-1)$ queries (each equally likely), one for each (i, j) pair with $i < j$. The query for the pair (i, j) (where i beat j without loss of generality) has two candidates (p, z), namely $(1, i)$ and $(0, j)$. Thus, if we show the ad corresponding to the winner (with $z = i$), we have $p = 1$, and the bid is 1, so we get value 1; if we show ad with $z = j$, we have $p = 0$, we get no value. It is then clear that the efficiency-maximizing ranking of the raw predictions z exactly corresponds to the solution to the MFAS problem.

Negative results for Q1, Q3, and Q4 in general We also show negative results for Q1, Q3, and Q4 in general.

For Q1, observe that in the NP-hardness construction when there is a perfect ranking, we observe a CTR of 1.0, and so the efficiency-maximizing prediction map cannot be self-calibrated. We can illustrate this directly with the following example. There are four ads, each given as (p, b, z) tuples:

q_1	q_2
A $(1.0, 2, z_1)$	C $(1.0, 2, z_2)$
B $(0.0, 2, z_2)$	D $(0.0, 1, z_1)$

We need $f(z_1) > f(z_2)$ in order to guarantee we show Ad A on q_1; we also need $f(z_2) > \frac{1}{2}f(z_1)$ in order to show Ad C on q_2. We will observe a 1.0 CTR on both z_1 and z_2 on any such efficiency maximizing f, but we are constrained to pick $f(z_2) < f(z_1) \leq 1$, and so no such f can be self-calibrated.

For Q3, we have already shown multiple fixed points in the single-query case. If we consider multiple queries, where each query has a single distinct raw prediction, we immediately arrive at a problem with exponentially many fixed points.

For Q4, it is straightforward to construct an example with cycles, but constructing one with no fixed point is a bit trickier. In particular, any time there is some prediction z where each query has at least one ad with prediction z, we can always find a fixed point by setting $f(z') = 0$ for $z' \neq z$ and $f(z) > 0$. The set of ads shown will be independent of the non-zero value $f(z)$, so we can set it equal to the observed CTR, achieving self-calibration (except in the degenerate case where all the ads with prediction z have zero CTR).

However, it is still possible to construct problems with no fixed points without resorting to such degeneracy, as the following example illustrates. Each query is equally likely, all the bids are 1, and the (p, z) ad tuples are:

q_1	q_2	q_1	q_2
A $(0.5, z_1)$	B $(0.6, z_2)$	C $(0.5, z_1)$	E $(0.2, z_2)$
		D $(0.6, z_2)$	F $(0.3, z_1)$

both	Prop E1 $\Longrightarrow$ Prop E2	(immediate)
ALL	Prop E2 $\Longleftrightarrow$ Prop SI	Thm 1
ALL	Prop E2 $\Longrightarrow$ nice	Thm 2
ALL	Prop E1 $\Longrightarrow$ nice	(from above)
ONE	Prop E1 $\Longrightarrow$ nice	Thm 3
both	Prop SI $\not\Longrightarrow$ Prop E1	Sec 5.2
ONE	Prop E2 $\not\Longrightarrow$ Prop SI	Sec 5.2
ONE	Prop SI $\not\Longrightarrow$ nice	Sec 5.2

Table 1: Relationships between problem properties. A "nice" problem instance is one where a self-calibrated efficiency-maximizing prediction map exists.

If $f(z_1) > f(z_2)$, then we show ads A,B, C, and F. In this case, we observe a CTR of $(0.5 + 0.5 + 0.3)/3 = 0.433$ for z_1, and 0.6 for z_2, so we cannot be self-calibrated. If $f(z_1) < f(z_2)$, we show ads A, B, D, and E, and observe a CTR of $(0.6 + 0.6 + 0.2)/3 = 0.467$ for z_2, and 0.5 for z_1, and so again we cannot be self-calibrated. Finally, if $f(z_1) = f(z_2)$, we always show A and B, and show the other ads half of the time. Thus, we observe a CTR of $(3/4)0.5+(1/4)0.3 = 0.45$ for z_1, and a CTR of $(3/4)0.6 + (1/4)0.2 = 0.5$ for z_2, and so again we cannot be well-calibrated. Thus, no self-calibrated f exists for this problem.

5 Sufficient Conditions

As the previous two sections show, without additional assumptions significant problems arise if one tries to achieve both calibration and auction efficiency. In this section, we introduce additional assumptions that are sufficient to guarantee nice prediction maps exist. Table 1 summarizes our results.

The intuition behind our results is a basic property of conditional probability. Calibration depends on the conditional expectation $\mathbb{E}[p \,|\, z]$. In general, selection changes the distribution this expectation is with respect to. But if selection is *only* a function of z, it does not change the conditional distribution of p given z, since the latter is already conditioned on z.

For example, suppose we have a single query, and that all bids are 1, so all selection decisions are functions of z. This means that $\mathbb{E}[p \mid z]$ does not change under selection, and thus defines an efficiency-maximizing self-calibrated prediction map. To extend this intuition to more realistic auctions, we need to make sure that the query and the bid do not add any information about p, so that selection does not change $\mathbb{E}[p \,|\, z]$ and the different $\mathbb{E}[p \,|\, z]$ for each query can be reconciled. We now state these properties formally:

Prop E1 For each z there exists a value $\bar{p}(z)$ such that for each query q with $\Pr_{\mathcal{C}}(q \,|\, z) > 0$, and for each b with $\Pr_{\mathcal{C}}(b \,|\, q, z) > 0$,

$$\mathbb{E}_{\mathcal{C}}[p \,|\, z, b, q] = \mathbb{E}_{\mathcal{C}}[p \,|\, z, q] = \mathbb{E}_{\mathcal{C}}[p \,|\, z] \equiv \bar{p}(z). \tag{1}$$

That is, in all cases the bid and query provide no more information than the raw prediction about average click-through rates.[4] For this assumption, the natural prediction map to consider is $f(z) = \bar{p}(z)$. □

Prop `E2` A weaker assumption is that

$$\mathbb{E}_{\mathcal{C}}[p \,|\, z, b] = \mathbb{E}_{\mathcal{C}}[p \,|\, z] \tag{2}$$

whenever both expectations are defined. This essentially marginalizes over queries, rather than holding simultaneously for all q. □

Prop `SI` A problem instance is **selection-invariant** if for all f, f', for any z where both $\mathbb{E}_f[p \,|\, z]$ and $\mathbb{E}_{f'}[p \,|\, z]$ are defined, we have

$$\mathbb{E}_f[p \,|\, z] = \mathbb{E}_{f'}[p \,|\, z]. \tag{3}$$

Selection invariance says that the observed CTR for a given raw prediction z is independent of the prediction map used for selection. Under this assumption, the natural calibration map to consider is $f^*(z) = \mathbb{E}_{f_z}[p \mid z]$, where f_z is any prediction map that shows some ads with raw prediction z. □

It is easy to show that Prop `E1` implies Prop `E2`.

A weak per-query variant of Prop `E1` is that, for all z, b, and q (when defined), $\mathbb{E}_{\mathcal{C}}[p \mid z, b, q] = \mathbb{E}_{\mathcal{C}}[p \mid z, q]$. We can dismiss this assumption as insufficient, as we can take the negative examples of Section 3 and re-state them where each candidate occurs on a distinct query, each equally likely. Thus, the above property holds trivially, but the pathological behaviors still occur.

5.1 Properties that Imply Nice Maps Exist

First, we show that under mechanism `ALL`, Prop `E2` and Prop `SI` are equivalent; we then show that Prop `E2` (and hence also Prop `SI`) imply a nice problem.

Theorem 1. *Under selection mechanism `ALL`, Prop `E2` is equivalent to Prop `SI` (selection invariance).*

Proof sketch. Suppose Prop `E2` holds. Selection mechanism `ALL` must show either all of the candidates with a given (z, b) combination, or none of them. Thus, for any f where $\Pr_f(z, b) > 0$, we must have

$$\mathbb{E}_f[p \,|\, z, b] = \mathbb{E}_{\mathcal{C}}[p \,|\, z, b]. \tag{4}$$

Then, for any f, assuming $\mathbb{E}_f[p \,|\, z]$ is defined,

$$\begin{aligned}
\mathbb{E}_f[p \,|\, z] &= \mathbb{E}_f[\mathbb{E}_f[p \,|\, z, b]] \\
&= \mathbb{E}_f[\mathbb{E}_{\mathcal{C}}[p \,|\, z, b]] && \text{Eq. (4)} \\
&= \mathbb{E}_f[\mathbb{E}_{\mathcal{C}}[p \,|\, z]] && \text{Prop } \texttt{E2} \\
&= \mathbb{E}_{\mathcal{C}}[p \,|\, z].
\end{aligned}$$

[4] Note that this does not hold under the NP-Hardness reduction for `ONE` in the previous section, as $\mathbb{E}_{\mathcal{C}}[p \,|\, z, q] \neq \mathbb{E}_{\mathcal{C}}[p \,|\, z]$.

For the other direction, suppose we have selection invariance (Prop SI). It is sufficient to consider a fixed raw prediction z (if there are multiple z, we can consider them independently). Also, we can assume candidates have distinct bids - if multiple candidates have the same bid and raw prediction, mechanism ALL treats them all the same, so we can just average over them.

Index the bids $(b_1, b_2, \dots)$ in decreasing order. Then, depending on the chosen $\hat{p} = f(z)$, we either show (when the appropriate queries occur) ad 1, or ads 1 and 2, etc. Prop SI says that no matter what $\hat{p}$ is, the average CTR of the ads we show is the same. Suppose that all the ads are on the same query. Then Prop SI implies $p_1 = \frac{1}{2}p_1 + \frac{1}{2}p_2$, so $p_1 = p_2$; $\frac{1}{2}(p_1 + p2) = \frac{1}{3}(p_1 + p_2 + p_3)$, so $p_1 = p_2 = p_3$; and so on. When the ads are on different queries, the weights in the above equalities change to reflect the query distribution, but are still all positive and sum to 1, so the same inductive reasoning holds. □

This result implies that under selection mechanism ALL, when Prop E2 holds the prediction map $f^*(z) = \mathbb{E}_{\mathcal{C}}[p \,|\, z]$ is self-calibrated. Next, we show this map is in fact also efficiency-maximizing:

Theorem 2. *Under selection mechanism ALL, Prop E2 implies f^* is efficiency maximizing, where $f^*(z) = \mathbb{E}_{\mathcal{C}}[p \,|\, z]$.*

Proof. Recall we need to show f^* maximizes

$$EV(f) = \sum_{i \in \mathcal{C}} \Pr^{\mathcal{Q}}(q_i) \Pr(i \,|\, q_i, f)(p_i b_i - 1).$$

Since selection decisions for one z value do not impact others, it suffices to consider a single z value. We can decompose the sum over $\mathcal{C}$ over the partition that associates all the ads that share a common bid and raw prediction. Let $B = \{i \,|\, b_i = b, z_i = z\} \subseteq \mathcal{C}$ be the element of this partition for (b, z). For a given $f(z) = \hat{p}$, either all the ads in B show (when their respective queries occur), or none of them do; thus, if we can show that f^* shows these ads if and only if they increase EV, we are done. The expected value per query of showing these ads is:

$$\sum_{i \in B} \Pr^{\mathcal{Q}}(q_i) \Pr(i \,|\, q_i, f)(p_i b_i - 1). \tag{5}$$

Since $\Pr(i \,|\, q_i, f) \in \{0, 1\}$ must be the same for all these ads, this quantity is non-negative if and only if $\sum_{i \in B} \Pr^{\mathcal{Q}}(q_i)(p_i b_i - 1) \geq 0$.

Recall $\Pr_{\mathcal{C}}(i) = \Pr^{\mathcal{Q}}(q_i)/C$ where $C = \sum_{i \in \mathcal{C}} \Pr^{\mathcal{Q}}(q_i)$. We have $\Pr_{\mathcal{C}}(i \wedge b \wedge z) = \Pr_{\mathcal{C}}(i)$ if $i \in B$, and 0 otherwise. Letting $C_B = \sum_{i \in B} \Pr^{\mathcal{Q}}(q_i)$, then $\Pr_{\mathcal{C}}(b \wedge z) = \frac{C_B}{C}$, and so

$$\Pr_{\mathcal{C}}(i \,|\, b, z) = \frac{\Pr_{\mathcal{C}}(i)}{C_B/C} = \frac{\Pr^{\mathcal{Q}}(q_i)/C}{C_B/C} = \frac{\Pr^{\mathcal{Q}}(q_i)}{C_B} \tag{6}$$

for $i \in B$, and 0 otherwise. Then,

$$\begin{aligned}
\mathbb{E}_{\mathcal{C}}[p \,|\, z] &= \mathbb{E}_{\mathcal{C}}[p \,|\, b, z] && \text{Prop } \texttt{E2} \\
&= \sum_{i \in \mathcal{C}} \Pr_{\mathcal{C}}(i \,|\, b, z) p_i \\
&= \frac{1}{C_B} \sum_{i \in B} \Pr^{\mathcal{Q}}(q_i) p_i. && \text{Eq. (6)}
\end{aligned}$$

Using this result, we have

$$\begin{aligned}
\sum_{i \in B} \Pr^{\mathcal{Q}}(q_i)(p_i b_i - 1) &= b \left(\sum_{i \in B} \Pr^{\mathcal{Q}}(q_i) p_i \right) - C_B) \\
&= C_B(b \mathbb{E}_{\mathcal{C}}[p \,|\, z] - 1).
\end{aligned}$$

This quantity is non-negative if and only if $bf^*(z) - 1 \geq 0$; since this is exactly the condition we use to decide whether or not to show the ads in B, we are done. □

It is not hard to directly prove that under selection mechanism `ALL`, Prop `SI` implies f^* is efficiency-maximizing: the idea is to consider again a single z, sort the ads by bid into blocks, and show by induction that each block has average CTR $f^*(z)$.

In Section 4 we saw that the problem of finding an efficiency-maximizing f is NP-hard under mechanism `ONE`, even under the assumption of a single bid. Under Prop `E1`, fortunately the situation is much easier:

Theorem 3. *Under selection mechanism `ONE`, if Prop `E1` holds then the prediction map f^* where $f^*(z) = \mathbb{E}_{\mathcal{C}}[p \,|\, z]$ is efficiency-maximizing and self-calibrated.*

Proof. For a query q, consider a partition $\mathfrak{B}^q$ of $\mathcal{C}(q)$ into sets of ads that share a common b and z, so the elements of the partition are

$$B^q_{b,z} = \{i \,|\, b_i = b, z_i = z, q_i = q\} \subseteq \mathcal{C}(q)$$

for each (b, z) pair.

All $i \in B$ for some B must share a common value $\Pr_f(i \,|\, q)$. We also use B as the event that some $i \in B$ shows; so for example $\Pr_f(B \,|\, q)$ is the probability that *some* ad from B shows. Under selection mechanism `ONE`, for each $i \in B$, we have $\Pr_f(i \,|\, B, q) = \frac{1}{|B|}$ (since ties are broken at random). Also,

$$\mathbb{E}_{\mathcal{C}}[p \,|\, b, z, q] = \frac{1}{|B^q_{b,z}|} \sum_{i \in B^q_{b,z}} p_i. \tag{7}$$

Recalling cost is zero under ONE, for any f,

$$\begin{aligned}
&\mathrm{EV}(f) \\
&= \sum_{q \in \mathcal{Q}} \Pr^{\mathcal{Q}}(q) \sum_{i \in C(q)} \Pr_f(i \,|\, q) p_i b_i \\
&= \sum_{q \in \mathcal{Q}} \Pr^{\mathcal{Q}}(q) \sum_{B^q_{b,z} \in \mathfrak{B}^q} \sum_{i \in B^q_{b,z}} \Pr_f(i \,|\, q) p_i b_i \\
&= \sum_{q \in \mathcal{Q}} \Pr^{\mathcal{Q}}(q) \sum_{B^q_{b,z} \in \mathfrak{B}^q} \Pr_f(B^q_{b,z} \,|\, q) \frac{1}{|B^q_{b,z}|} \sum_{i \in B^q_{b,z}} p_i b
\end{aligned}$$

and using Eq. (7),

$$\begin{aligned}
&= \sum_{q \in \mathcal{Q}} \Pr^{\mathcal{Q}}(q) \sum_{B^q_{b,z} \in \mathfrak{B}^q} \Pr_f(B^q_{b,z} \,|\, q) b \, \mathbb{E}_{\mathcal{C}}[p \,|\, b, z, q] \\
&\leq \sum_{q \in \mathcal{Q}} \Pr^{\mathcal{Q}}(q) \max_{B^q_{b,z} \in \mathfrak{B}^q} b \, \mathbb{E}_{\mathcal{C}}[p \,|\, b, z, q].
\end{aligned}$$

Thus, it is sufficient to show that selecting ads using f^* produces the expected value in the last line of the above inequality. For each query, we rank the ads using $b \cdot f^*(z) = b \, \mathbb{E}_{\mathcal{C}}[p \,|\, b, z, q]$, and so this is exactly the expected value that f^* obtains.

To see that f^* is self-calibrated, observe that when $\Pr_f(z, b, q) > 0$,

$$\mathbb{E}_f[p \,|\, z, b, q] = \mathbb{E}_{\mathcal{C}}[p \,|\, z, b, q] = f^*(z),$$

and so

$$\mathbb{E}_f[p \,|\, z] = \sum_{b,q} \Pr_f(b, q \,|\, z) \mathbb{E}_f[p \,|\, z, b, q] = f^*(z).$$

□ □

5.2 Negative Results

We show several negative results relating to the assumptions considered in the previous section.

ONE and ALL: Prop SI does not imply Prop E1 Consider an example with two queries, each equally likely. Each query has two candidates, given as the following (p, b) tuples (they all share a common z):

q_1	q_2
A (0.1, 1)	C (0.1, 2)
B (0.2, 2)	D (0.2, 1)

Because of the symmetry between these queries, under any f (and either selection mechanism), ad A must show with the same probability as ad D, as must ads B and C. Thus, for any f, $\mathbb{E}_f[p \mid b = 1, z] = 0.15$, and similarly $\mathbb{E}_f[p \mid b = 2, z] = 0.15$. Thus, selection invariance holds, as does Prop `E2`. However, $\mathbb{E}_{\mathcal{C}}[p \,|\, z, b = 1, q_1] = 0.1 \neq \mathbb{E}_{\mathcal{C}}[p \,|\, z, q_1] = 0.15$.

`ONE`: Prop `E2` does not imply Prop `SI` Consider the example, with two equally likely queries, and two distinct raw predictions:

q_1	q_2
A $(0.2, 2, z_1)$	C $(0.1, 2, z_1)$
B $(0.1, 1, z_1)$	D $(0.2, 1, z_1)$
E $(1.0, 9, z_2)$	

Note that $\mathbb{E}_{\mathcal{C}}[p \mid z_1, b = 1] = \mathbb{E}_{\mathcal{C}}[p \mid z_1, b = 2] = 0.15$. However, if we consider two prediction maps $f(z_1) = 0.5, f(z_2) = 1$ and $f'(z_1) = 1, f'(z_2) = 0$, under selection mechanism `ONE`, we have $\mathbb{E}_f[p \,|\, z_1] = 0.1$, but $\mathbb{E}_{f'}[p \,|\, z_1] = 0.15$.

`ONE`: Prop `SI` does not imply a nice problem We have four queries, each equally likely; the bids for the ads on q_3 and q_4 are defined in terms of some small $\epsilon > 0$, with (p, b, z) tuples:

q_1	q_2	q_3	q_4
A $(1, 2, z_1)$	C $(1, 2, z_2)$	A' $(0, 2\epsilon, z_1)$	C' $(0, 2\epsilon, z_2)$
B $(0, 2, z_2)$	D $(0, 1, z_1)$	B' $(1, 2\epsilon, z_2)$	D' $(1, 1\epsilon, z_1)$

Note that q_3 and q_4 mirror q_1 and q_2, except that the bids are scaled by ϵ, and the CTRs are reversed. Under any f, ads A and A' show with the same probability, as do B and B', and the other two pairs. Thus, under selection by any f, we have $\mathbb{E}_f[p \,|\, z_1] = \mathbb{E}_f[p \,|\, z_2] = 0.5$ whenever the expectation is defined, and so Prop `SI` holds. However, as $\epsilon \to 0$, only q_1 and q_2 have any impact on efficiency. Thus, as before we have constraints on the optimal solution that $f(z_1) > f(z_2) > \frac{1}{2} f(z_1)$. Thus, the prediction map f^* with $f^*(z_1) = 0.5$ and $f^*(z_2) = 0.5$ is not efficiency-maximizing, as it only shows ad A on q_1 only half the time.

6 Discussion and Future Work

Our sufficient conditions are quite strong, but not unrealistic. They require that the bid and query not add any information about the CTR, conditional on the raw prediction. CTR estimation systems normally use queries as features (e.g., [13]), so it is reasonable to hope that the query does not add extra information. Bids are set by advertisers for query-ad pairs, which are already used by CTR estimation systems, so any systematic patterns in bids are likely to be accounted for. Since advertisers have much less information than the auctioneer, it seems unlikely that they can add extra information about CTRs through fine-grained

bid manipulation. We can test if our sufficient conditions hold by running randomization experiments that change the mix of ads shown.

Since randomized predictions cannot in general lead to maximum efficiency, it is natural to first consider deterministic prediction maps. Nevertheless, given the negative results in the current work, it would be interesting to also study randomized calibration strategies that provide calibration guarantees without needing IID assumptions. Then the natural question becomes: how much efficiency is lost by using a randomized calibration strategy, versus using a deterministic efficiency-maximizing prediction map that is not self-calibrated.

References

[1] Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

[2] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06. ACM, 2006.

[3] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, ICML 2006, 2006.

[4] Nicolò Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006. ISBN 0521841089.

[5] Ira Cohen and Moises Goldszmidt. Properties and benefits of calibrated classifiers. In *8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*. Springer, 2004.

[6] A. Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 1982.

[7] Morris H. DeGroot and Stephen E. Fienberg. The comparison and evaluation of forecasters. *The Statistician*, 32, 1983.

[8] Benjamin Edelman, Michael Ostrovsky, and Michael Schwarz. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American Economic Review*, 97(1):242–259, March 2007.

[9] Dean Foster and Rakesh Vohra. Asymptotic calibration. *Biometrika*, 85: 379–390, 1996.

[10] Dean P. Foster. A proof of calibration via blackwell's approachability theorem. *Games and Economic Behavior*, 29(1-2), October 1999.

[11] Tilmann Gneiting and Adrian Es Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 2007.

[12] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2), 2007.

[13] Thore Graepel, Joaquin Quiñonero Candela, Thomas Borchert, and Ralf Herbrich. Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine. In *ICML*, 2010.

[14] Robert Kleinberg, Alexandru Niculescu-Mizil, and Yogeshwer Sharma. Regret bounds for sleeping experts and bandits. *Machine Learning*, 80.

[15] Niculescu-Mizil, Alexandru, and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, ICML 2005, 2005.

[16] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, ICML '05. ACM, 2005.

[17] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*. MIT Press, 1999.

[18] Roopesh Ranjan and Tilmann Gneiting. Combining probability forecasts. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1), 2010.

[19] Hal Ronald Varian. Position auctions. *International Journal of Industrial Organization*, 25(6), 2007.

[20] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02. ACM, 2002.